DOCUMENT RESUME

ED 471 082 TM 034 587

AUTHOR	Pearson, P. David; Calfee, Robert; Webb, Patricia L. Walker; Fleischer, Steve
TITLE	The Role of Performance-Based Assessments in Large-Scale Accountability Systems: Lessons Learned from the Inside. Technical Guidelines for Performance Assessment.
INSTITUTION	Council of Chief State School Officers, Washington, DC.
SPONS AGENCY	Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE	2002-00-00
NOTE	44p.; Produced by the State Collaborative on Assessment and Student Standards.
CONTRACT	R279A50006
PUB TYPE	Reports - Research (143)
EDRS PRICE	EDRS Price MF01 Plus Postage. PC Not Available from EDRS
DESCRIPTORS	Accountability; Case Studies; *Large Scale Assessment; *Performance Based Assessment; Professional Development; State Programs; *Test Construction; *Testing Programs
IDENTIFIERS	Large Scale Programs

ABSTRACT

In 1996, a subcommittee of the State Collaborative on Assessment and Student Standards commissioned a study of the use of performance-based assessments in large-scale accountability systems. The idea was to look into current state assessment work on performance-based assessments to see what has been learned, but not widely reported, by those who had been at work I the trenches. Five states were included in the case studies: Maryland, Kentucky, Oregon, Washington, and Delaware. In each state, interviews were held with key participants and stakeholders in the statewide assessment effort. In addition to the interviews, data were obtained from document reviews. The lessons learned are summarized in these points: (1) pay attention to the technical qualities; (2) successful assessment is more than just technical; (3) building support is essential; (4) establish system accountability before student accountability; (5) nurture professional development about and through the assessments; (6) it is most than just professional development; (7) if possible do not send a test out to do a framework's job; (8) time may be a luxury but it buys a lot; and (9) the most important asset of an assessment system is the people, especially the people in the middle. An appendix contains the interview protocol. (Contains 11 endnotes and 31 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 471 082

SCASS
State Collaborative on Assessment and Student Standards

The Role of Performance-Based Assessments in Large-Scale Accountability Systems:

Lessons Learned from the Inside

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

B. Butterbaugh

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

P. David Pearson
Robert Calfee
Patricia L. Walker Webb
Steve Fleischer

TM034587

SCASS TGPA
Technical Guidelines for Performance Assessment

A Publication of the Council of Chief State School Officers

BEST COPY AVAILABLE

The Council of Chief State School Officers (CCSSO) is a nationwide, nonprofit organization composed of the public officials who head departments of elementary and secondary education in states, the District of Columbia, the Department of Defense Education Activity, and five extra-state jurisdictions. CCSSO seeks its members' consensus on major education issues and expresses their views to civic and professional organizations, to federal agencies, to Congress, and to the public. Through its structure of standing committees and special task forces, the Council responds to a broad range of concerns about education and provides leadership on major education issues. Because the Council represents each state's chief education administrator, it has access to the educational and governmental establishment in each state and to the national influence that accompanies this unique position. CCSSO forms coalitions with many other education organizations and is able to provide leadership for a variety of policy concerns that affect elementary and secondary education. Thus, CCSSO members are able to act cooperatively on matters vital to the education of America's young people.

The State Education Assessment Center is a permanent, central part of the Council of Chief State School Officers. The Center was established through a resolution by the membership of CCSSO in 1984. This report is sponsored by the Assessment Center's State Collaborative on Assessment and Student Standards (SCASS), Technical Guidelines for Performance Assessment (TGPA) consortium. The SCASS TGPA works with researchers to design and implement practical and timely research on large-scale performance assessment. This research provides information useful in designing state assessment and accountability programs so that they yield results that can be used to improve student learning.

COUNCIL OF CHIEF STATE SCHOOL OFFICERS

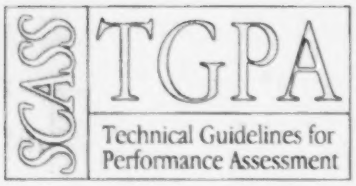
G. Thomas Houlihan
Executive Director

Wayne N. Martin
Director
State Education Assessment Center

John F. Olson
Director of Assessments
State Education Assessment Center

Phoebe C. Winter
Project Director, Technical Guidelines for Performance Assessment
State Collaborative on Assessment and Student Standards

Copyright © 2002 by the Council of Chief State School Officers, Washington, DC



The Role of Performance-Based Assessments in Large-Scale Accountability Systems:

Lessons Learned from the Inside

P. David Pearson
Robert Calfee
Patricia L. Walker Webb
Steve Fleischer

Council of Chief State School Officers
One Massachusetts Avenue, N.W.
Washington, DC 20001

This report was prepared for submission under contract with the Council of Chief State School Officers and funded by the United States Department of Education, Office of Educational Research and Improvement, Grant No. R279A50006. The views and opinions expressed in this report are not necessarily those of the United States Department of Education, the Council of Chief State School Officers, the State Collaborative on Assessment and Student Standards, or the states participating in the study.

ACKNOWLEDGMENTS

This project was conducted with the assistance of several members of the State Collaborative on Assessment and Student Standards (SCASS), Technical Guidelines for Performance Assessment (TGPA), who helped structure the study, assisted in the data collection, and reviewed draft versions of the report:

STUDY GROUP ON PERFORMANCE-BASED ASSESSMENTS AND ACCOUNTABILITY

Elaine Grainger, Chair	Vermont Department of Education*
Peter Behuniak	Connecticut Department of Education
Ellen Hedlund	Rhode Island Department of Education
Barbara Kapinus	Council of Chief State School Officers*
James Masters	Pennsylvania Department of Education
Sue Rigney	Kentucky Department of Education*
Steve Slater	Oregon Department of Education
Phoebe Winter	Council of Chief State School Officers
Shu-Jing Yen	Maryland Department of Education

* At time of development.

EXECUTIVE SUMMARY

In 1996, a subcommittee of the State Collaborative on Assessment and Student Standards (SCASS), Technical Guidelines for Performance Assessment (TGPA), all individuals actively involved in statewide assessment, commissioned this study of the use of performance-based assessments in large-scale accountability systems. The idea was to get inside current state assessment work on performance-based assessment to shed some light on what had been learned, but not widely reported, by those who had been at work in the trenches, making performance assessment a viable part of their state assessments. This work would represent an insider, or emic, perspective on the process rather than the more common external, or etic, perspective of the sort documented in independent evaluations of state assessment efforts. The expectation was that what was learned by studying this work from the perspective of those directly involved in the assessment effort would be useful to state assessment personnel, as well as those who bear responsibility for assessment in large districts who are either considering or on the verge of implementing performance assessments.

Five sites were included in the case studies: Maryland, Kentucky, Oregon, Washington, and Delaware. Within each site, interviews were held with key participants and stakeholders in the statewide assessment effort. Participants included individuals working in the state department of education and district and school personnel. In addition to interviews, we conducted extensive reviews of documents related to the policy, design, and implementation of the state assessment systems.

LESSONS LEARNED

Our conclusions are organized in a set of lessons we learned as we collected and analyzed the data for this study.

1. Pay Attention to the Technical Qualities.

While all states have attended to technical issues, in Maryland they were elevated to center stage from the outset. In retrospect and if longevity is any index of wisdom, this turns out to have been a wise decision; the Maryland assessment is still with us, albeit in gradually evolving form, while other bolder efforts are now historical artifacts.

2. But It Is More than Just Technical.

Technical quality is important, but it does not guarantee a good or a credible assessment. The technical qualities are necessary, but not sufficient, to achieve broad acceptance. For broad acceptance, communication appears to be the key. In states where proponents have "taken the test on the road" to bring educators, policymakers, and the general public up to speed, understanding has led to broader support, which brings us to our next lesson.

3. Building Support Is Essential.

Individuals in all five states talked extensively about building a base of support among all key constituencies, especially those with histories of resistance or those who are vulnerable to opposition campaigns. Building support involves a combination of

politics, information, and communication. Developing materials that can be shared broadly with the public is essential. Finally, there is no substitute for getting the word out in public and professional meetings where the performance-based assessments (PBAs) can be explained and where concerns can be expressed and responded to directly.

4. *Establish System Accountability before Student Accountability.*

If system-level accountability precedes individual student accountability, it increases the likelihood that the resources—professional development, curricular practices, a bank of local on-demand tasks, and the like—needed to support students in their quest for higher performance on more challenging tasks will be in place before consequences set in. As these multiple-level (students, schools, and/or districts) accountability systems play themselves out, legal issues are likely to emerge, as they have in virtually all high-stakes enterprises. While unlikely, it is not implausible to expect rulings that forbid the imposition of individual-level accountability before school and/or district accountability systems have been established and the necessary resources (i.e., the curriculum, materials, and professional development) have been put into place.

5. *Nurture Professional Development about and <u>through</u> the Assessments.*

This is a composite lesson built from the experience of each state. Teacher involvement in every step of the assessment process, including task development, rubric development, scoring, and selecting anchors, not only builds ownership and commitment to the process (not to mention its impact on the quality of the assessments themselves), it also builds local capacity; that is, it helps districts and schools to develop their own PBA tasks to use for school- and classroom-based decision-making and to make sure that teachers and students are prepared for the PBA component of the state assessment. Closer partnerships between teacher education and the public schools they serve would enable preservice teachers to get a firm grounding in the use of a wide range of assessment tools, including performance assessments, curriculum-based assessments, and informal classroom assessment.

6. *But It Is More than <u>Just</u> Professional Development.*

Professional development can provide the knowledge base, and it can even build ownership and commitment among those who participate actively in the assessment process. It can even go a long way to changing deeply held beliefs about the role teachers can play in improving student learning. However, it is unclear whether professional development can change teacher beliefs about the right and responsibility of the state (or the district or any policy body) to hold teachers and schools accountable for student performance. Belief in accountability systems, accompanied by a personal commitment to accepting the responsibility (of improving student achievement) that comes with it, is more a moral and ethical (and political) issue than a cognitive and technical matter. Strong, powerful communication, instantiated as persuasion and moral suasion, must accompany technical messages. Persuasion will help, but nothing succeeds like success—when teachers see dramatic changes in student performance and in students' willingness to accept responsibility for their own growth and achievement.

7. *If Possible, Do not Send a Test out to Do a Framework's Job.*

When states are able to build assessments alongside the standards or after the standards have been developed and disseminated, the assessments themselves are

likely to encounter less resistance than if they are developed before the standards, perhaps because constituents are better able to focus their concerns on the appropriate reform tool. Even more important is the contextualizing role that standards and curriculum frameworks can provide for assessment items and tasks. If an assessment artifact can be tied directly to a broadly accepted framework or set of standards, it does not have to stand on its own; the framework provides the assessment artifact with an aura of legitimacy that it might not possess on its own.

8. Time May Be a Luxury but It Buys a Lot.

Implicit in many of these lessons is the assumption that time can relieve many of the pressures and sources of resistance that PBAs experience when they enter the policy arena. Time gives us space for a more deliberate development process; provides for a more elaborate process of communication and public outreach; ensures that assessments will be accompanied by other policy tools, such as frameworks and standards; and acknowledges the reality that it takes more than a single school year for teachers, other educators, and the general public to acknowledge and accept the fact that new tools and practices are here to stay. And time, if accompanied by appropriate resources, gives states, districts, and schools a fighting chance to implement the professional development and to develop the local counterparts of the state assessment tools that make the system work more smoothly and give teachers and students the opportunities they need to put their best foot forward when the assessment becomes a reality.

9. The Most Important Asset of an Assessment System Is the People, Especially the People in the Middle.

Much of the success achieved thus far in these five states must be attributed to the individuals who, at the state, regional, district, and local levels, turn plans for PBAs into realities that play themselves out in schools and classrooms. These are what one of our informants called the people in the middle, those state-, district-, and local-level personnel who turn the mandates imposed by the policymakers into realities that can be tolerated (if not embraced) by the constituents of the systems. Our conclusion about the important role played by these individuals came not so much from what was said in the interviews as from the sense of dedication and commitment we took away from the interviews. There was a strong sense that this sort of effort has the potential to improve the daily curricular life of students and teachers and to provide parents with better evidence about how their children and our schools are doing. This respect for and reliance on committed educators is probably what was behind the statement, which we heard in one form or another in every site, that, "It is technical, for sure, but it is more than just technical."

CONTENTS

ACKNOWLEDGMENTS iii

EXECUTIVE SUMMARY v

Lessons Learned v

INTRODUCTION 1

METHOD 3

Selecting Sites 3
Selecting Participants 3
The Interview Protocol 4
Other Data Sources 4
Data Analysis Procedures 4
Case studies 4
Cross-site analysis 5
A cautionary note 5

RESULTS 7

A Brief Overview of the Cases 7
Delaware 7
Kentucky 7
Maryland 9
Oregon 9
Washington 10
Looking across Cases 10
Context and Purpose 10
Development and Scoring 13
Reporting 14
Technical Issues 14
Relationship to Curriculum 15
Assessment and State Curriculum Efforts 15
Impact at the local level 16
Professional Development 19
Positive effects of PBAs 19
Concerns about professional development 20
Teacher preparation 20
Consequences 20
Rewards and sanctions 20
Resistance and support 21

LESSONS LEARNED IN THE CROSS-SITE ANALYSIS 23

1. Pay Attention to the Technical Qualities. 23
2. But It Is More than *Just* Technical. 23
3. Building Support Is Essential 24
4. Establish System Accountability before Student Accountability 24
5. Nurture Professional Development about and *through* the Assessments. 24
6. But It Is More than *Just* Professional Development. 25
7. If Possible, Do not Send a Test out to Do a Framework's Job. 26
8. Time May Be a Luxury but It Buys a Lot. 26
9. The Most Important Asset of an Assessment System Is the People, Especially the People in the Middle 27

REFERENCES 29

RELEVANT STATE-LEVEL RESOURCES 31

APPENDIX 33

Interview Protocol 33

The Purpose of this Project: 33

Consent Statement: 33

Opening Questions: 33

The Purpose Of The Assessment 34

ENDNOTES 39

INTRODUCTION

It has been over a decade since the performance assessment movement—at least as a serious effort on the part of classroom teachers, school and district assessment leaders, and state assessment policy makers—has had much currency in educational assessment. It has a much longer tradition in other parts of the world where performance tests have been used for decades, if not centuries, as gatekeeping devices for key points of passage—such as for placement in academic versus vocational tracks in secondary school and for college admission. Even in the United States there have been domains in which performance assessments have been the rule rather than the exception—athletics, art, music, speech, architecture, and, to a lesser extent, law, medicine, and accounting. In those fields, individuals are judged to have achieved certain standards or levels of accomplishment on the basis of scores on a performance rather a multiple choice test. And, there have been small pockets of influence for many years in a broad array of school subjects. For example, portfolios have been used for over 25 years as the basis for graduation from Walden III High School in Racine, Wisconsin (Mabry, 1992), and for over 15 years at Central Park East in New York City (Meier, 1995). Additionally, the Advanced Placement examination for art, administered by the College Board and Educational Testing Service, has required and evaluated portfolio submissions for several decades.

Nonetheless, it is only in the past decade that the formal use of performance assessment has generated widespread interest among teachers, policy makers, and psychometricians (Linn, 2000; Linn, Baker, & Dunbar, 1991; Valencia, Hiebert, & Afflerbach, 1994). This newfound scholarly interest has been accompanied by increasing popularity of the movement in our schools. Performance assessment has been championed by several special interests, ranging from those who think that multiple-choice tests are a major obstacle to progress in our schools (e.g., Fairtest, 1995; Wiggins, 1993) to those who believe that assessment can lead the charge in standards-based reform (e.g., Resnick & Resnick, 1992; Simmons & Resnick, 1993) to those who, while suspicious about the political uses of most, if not all assessments, see some merit in portfolios because of their focus on teacher empowerment and/or student responsibility (Hansen, 1992; Tierney, Carter, & Desai, 1991).

Performance-based assessments have also penetrated the large-scale assessment enterprise of states and large school districts in the early 1990s, but even there, they were not an entirely new phenomenon. By 1990, several states had a two-decade track record in the use of writing assessment, where direct writing assessments had replaced multiple-choice editing tasks as indicators of student writing achievement.. But it was not until the early 1990s that the potential role of performance assessments in large-scale assessment was seriously considered for all content areas. They entered the large-scale assessment landscape because they had the support of key constituents.

a) Business leaders (Tucker, 1988) saw in them a transparent relationship between performance tasks and desirable employee attributes (e.g., writing memos or using math skills to solve practical problems).

b) Teachers viewed them as more closely connected to their everyday classroom curriculum (Brenner, Pearson, Boyd, & Prico, 1996; Hoffman, et al., 1996).

BEST COPY AVAILABLE

c) Educational reformers, as we already indicated, saw them as a tool to leverage curriculum change, particularly toward a more problem-oriented curriculum (Resnick & Resnick, 1992).

d) Some members of the assessment community embraced the challenge of developing psychometric tools that could be used to gauge the reliability, validity, utility, and integrity of these nonconventional assessments (Linn, Baker, & Dunbar, 1991).

In 1996, a subcommittee of the CCSSO State Collaborative on Assessment and Student Standards (SCASS), Technical Guidelines for Performance Assessment (TGPA), all individuals actively involved in statewide assessment, commissioned this study of the use of performance-based assessments in large-scale accountability systems. The goal was to get inside current state assessment work on performance-based assessment to shed some light on what had been learned, but not widely reported, by those who had been at work in the trenches, making performance assessment a viable part of their state assessments. This work would represent an insider, or emic, perspective on the process rather than the more common external, or etic, perspective of the sort documented in independent evaluations of state assessment efforts. The expectation was that what was learned by studying this work from the perspective of those directly involved in the assessment effort would be useful to state assessment personnel, as well as to those who bear responsibility for assessment in large districts who are either considering or on the verge of implementing performance assessments. It is this expectation that has guided the work reported in this document.

METHOD

SELECTING SITES

Initially we selected four state sites for our case studies, adding a fifth, Delaware, when it became clear that it was in our mutual interest (both ours as a research team and Delaware's as a state) to collaborate in this venture. We wanted two sites with extensive experience in using performance-based assessments as a part of their statewide assessment and accountability system. As a balance, we wanted two sites that were on the cusp (i.e., still working toward full implementation of their system). Within the veteran category, we wanted one site in which the system was and had been fairly stable for some time and a second site in which the system was in transition. Our veteran sites were Maryland, in the stable category, and Kentucky, in the transition category. Originally, we selected Oregon and Washington as sites that were, when the effort began, working toward full implementation. Delaware, when it came into the study, gave us a site that was a bit ahead of Oregon and Washington in entering the full implementation stage of its system, which included a substantial PBA component. Within the overall set, we also wanted to meet some additional criteria: We wanted at least one state where the process had been fairly quiet and another where it had been decidedly public and vocal. We wanted one state where implementation was imminent (Delaware met this criterion) and at least one with at least a few years of breathing room ahead of it (Washington met this criterion). Finally, we wanted a contrast between a state where the timeline from conceptualization to implementation was ambitious (within a year or two) and another where it had been more protracted and deliberate (four years or more).

SELECTING PARTICIPANTS

Within each site, we set out to hold hour-long interviews with key participants and stakeholders in the statewide assessment effort. Interviewees included the following:

a) individuals working within the assessment division of the state department of education (e.g., perhaps the director and a few individuals directly responsible for particular content area assessments);

b) individuals who had oversight responsibility for the assessment effort (e.g., perhaps a deputy superintendent of education);

c) individuals within the department of education whose work might be expected to be influenced by the assessment (e.g., curriculum division personnel); and

d) various school personnel (e.g., district-level superintendents, assessment coordinators, or curriculum specialists, school principals, and/or classroom teachers).

The list of participants was negotiated separately for each state and tailored to our mutual interests and needs. In each site, six to ten individuals participated in the interviews.

BEST COPY AVAILABLE

The Interview Protocol

The full, generic interview protocol is provided in the Appendix. Basically, we developed a general set of questions that could be adapted to the interests and expected knowledge base of each type of participant. Further, not all questions were asked of each participant because they were not all equally appropriate. The guiding questions were derived from a concept paper about the role of PBAs in large-scale assessment currently in press by two of the authors of this study (Calfee, Pearson, & Rodriguez, in press). The questions fell into these broad categories:

a) the context and purpose of the assessment;

b) the relationship of the assessment to state curricular efforts;

c) the development process;

d) administration and scoring;

e) reporting systems; and

f) other issues.

The protocol development process was iterative. We developed sample questions, revised them as a research team, tried them out with critical colleagues who served in roles similar to the target population, reviewed them with the advisory committee from the TGPA, and repeated the process until we were all satisfied that the questions served our collective purposes. We conducted the interviews in site visits (lasting from three to five days), working with a local contact to schedule interviews for the members of our team collecting the data.

Other Data Sources

In addition to the transcripts of the interviews, we conducted extensive document reviews in two stages. In stage one, we asked participating states to share with us what we called their "top of the shelf" documents, those that were readily available for distribution to us, other professionals, and even the general public, either in print form or on an Internet site. We reviewed these documents to get a general sense of the history and trajectory of each assessment program. In stage two, which occurred in conjunction with the interviews, we sought less readily accessible documents, ones that we thought might give us more insight into the issues that were unearthed during the interview process. As it turned out, stage two searches turned up not so much inaccessible documents as documents that had been made public between the time of our top of the shelf analysis and the interviews. For example, almost every state had revised its major position papers on standards and assessment between stages one and two. Additionally, in all states, new legislation had been passed between 1997 and 1999. These bills, along with documents outlining statewide implementation, were also consulted.[1]

Data Analysis Procedures

Case studies

After transcribing the interviews, we used open coding techniques (searching for regularities, patterns, and insights) to create a case study for each of the five sites.

Our goal in each of the case studies was to provide an account of the purpose, development, and use of PBAs in that state that was authentic and insightful, that captured what was unique and informative about the PBA experience in that state. Once we were satisfied that the case studies provided as rich an account as we could muster, we shared them with state personnel to make sure that our characterization of their experience was consistent with what they had seen in their states and had said in the interviews with us.

Cross-site analysis

We then undertook the cross-site analysis that undergirds the current paper. We examined the case studies, searching again for regularities, patterns, and insights, this time across rather than within sites, and we also looked for differences and discontinuities across sites, on the grounds that what distinguished one state from the others might be as useful to other players in the PBA arena as were the consistencies. We made charts in which we explicitly compared the states on each of the major categories that were used to generate our interview protocol. Finally, we met regularly, in person and by telephone, as an analysis team to share insights and observations. The entire process was highly iterative, with constant journeys across the boundary between the data and our interpretations. We used data to generate hypotheses, and we returned to the data to evaluate the validity and generalizability of those tentative hypotheses, revising where necessary until we felt that the claims we were making were warranted by the evidence presented to support them. Finally, we presented a draft to individuals within each state and to the SCASS TGPA committee overseeing this effort as a way of testing the validity of the procedures we used. Based on feedback from that committee, we returned to the data and our interpretations to reassess and revise the claims we had made and the data available to warrant them.

A cautionary note

In the world of statewide assessment, things happen fast, and they change even faster. Because assessment is a matter of concern to many players and many constituencies, the policies operative at time A are likely to change by time B. One of the challenges with a study like ours is that it is like a snapshot taken of a moving train. We collected most of our data in the fall of 1998 through the summer of 1999, with follow-up phone calls to check on the accuracy and currency of our findings through the spring of 2000. Things have changed in each of these states, and things are likely to change more in the future. Even so, what is important in each case, as the results and the lessons learned will reveal, is the process by which the entire effort is carried out and the relationships among the individuals who are charged with the responsibility of making the system work. Thus, even accounts of systems that have since been modified are likely to be useful to us as we contemplate new and revised efforts in our own situations.

A Brief Overview of the Cases

We begin with a brief introduction to the PBA experience in each of the five case study states as a way of ensuring an adequate contextual background for the cross-site analysis that follows.

Delaware

The current Delaware program was initiated as an interim effort in 1994 after a determination was made that a performance-based element must be added to the statewide assessment program in order to assess all of the statewide goals and standards. The assessment is closely aligned to a set of statewide content standards that have been in place since 1995. PBA components have been available in reading, writing, and math since 1998, with science and social studies slated to go on-line in 1999–2000. A mixed model of assessment (some multiple-choice and some performance items) is used, and separate scores will be reported for each component; additionally, a standards-based reporting system will combine data from the two components to determine whether students have met particular standards. One unique feature of the Delaware system is the plan for on-line reporting of school- and district-level results, which makes possible its unusual combination of normative and standards-based reporting. By taking advantage of standards, expressed as particular cut scores on the tests, and precise normative data reports (e.g., comparing school scores with the state average or the average for demographically similar schools), school officials can build tailored reports that serve their own interests and needs.

The standards-based reporting will carry consequences in Delaware. For an individual, the standards-based reporting reveals whether an individual has met particular standards; for a school or district, the percentage of individuals within a school or district who have met standards is reported. The primary purpose of the assessment is to promote accountability for student achievement at all levels within the educational system and to encourage curricular reform in schools with low performance. In 2001, the accountability system begins in earnest, with rewards and sanctions prominently featured, most notably in the form of accreditation status for schools and scholarships for students. Our interviews led us to conclude that the overall reaction to the pilot assessments and to the accountability system more generally has been largely positive. Support is strong in most constituencies—the business community, teachers, administrators, state and local policymakers, and the general public, but, as officials readily admit, the imposition of high stakes in 2001 may affect support, at least among some of those constituent groups. Teachers are involved, either in an active or advisory role, in every phase of the process, but their most active role is in the development of the performance tasks for the PBA component.

Kentucky

One of the genuine trailblazing experiences in PBA, the Kentucky initiative began with the Kentucky Education Reform Act of 1990. The Reform Act was a response to a court decision over a lawsuit brought in 1985 challenging equity in school funding. The court ruling mandated that the state create a "new and constitutional system of public education" (Kentucky Department of Education, 1999). A task

force appointed in 1989 developed an entirely new whole-school improvement accountability system that held schools accountable for continuous improvement in student performance. The Kentucky Instructional Results Information System (KIRIS) was launched in 1991 to provide the data to drive the accountability system. KIRIS was last given in 1998, after which it was replaced by the Commonwealth Accountability Testing System (CATS), which, according to our participants, repositions the various components that comprised KIRIS. KIRIS consisted of open-response questions, such as essay questions and writing prompts; performance events; and portfolios (in writing and math) to demonstrate a student's best work from the school year. All students in grades 4, 8, and 12[2] were tested in seven areas: reading, writing, social science, science, mathematics, arts and humanities, and practical living/vocational studies. KIRIS was designed as a school accountability system, with some provision for district-level reporting and consequences; while individual student scores were (and are) reported, there are few consequences, positive or negative, for students.

In addition to the extensive review process that items and tasks underwent before they became part of KIRIS (bias review, content area advisory committees, etc.), the entire range of PBAs (portfolios, performance-tasks, open-ended test items) in Kentucky have been carefully scrutinized by both internal and external groups. Teachers have been involved in several aspects. Regarding the tasks themselves, teachers have been involved in task development and portfolio evaluation, but performance task scoring is handled by external contractors. Teachers are also involved in advisory committees, but perhaps their most important role is providing, by statute, the local professional development required to roll out the portfolio system in writing.[3]

With KIRIS, schools were provided rewards for exceeding their individually determined achievement goals (based upon a complex formula that takes into account their initial levels of performance and other contextual variables such as attendance and dropout rates). When achievement fell far short of those goals, the state was authorized to bring in both financial resources and extra personnel (in the form of an external school improvement facilitator).

The KIRIS results were consistently subjected to stringent quality control measures, including "extensive monitoring of scoring consistency and multiple independent replications of results for equating" (Kentucky Department of Education, 1999). Even though improvements in these quality control indices occurred over time, the improvement could not prevent KIRIS from being replaced by CATS in 1999. While the system has experienced tremendous support, especially from the business community and policy makers, it has also experienced resistance from teachers (some pockets of teachers have lamented the erosion of professional prerogative), students, and parents (who see little value taking up good instructional time for an activity that has no potentially positive consequences for individual students). The early accusations of low marks on technical characteristics (e.g., inadequate equating procedures and poor interjudge reliability on the portfolios in writing and math—see Koretz, 1996, and Koretz & Barron, 1998) made KIRIS especially vulnerable to opponents, even those who opposed it for conceptual or political rather than technical reasons. The recent transformation of KIRIS into CATS was undertaken to clarify the nature and function of the assessments and the accountability system; these are now explicitly stated in the legislation rather than a matter of implementing the legislative mandate. According to our participants, key components of KIRIS (performance tasks, the writing portfolio, and an alternate portfolio for students with disabilities) are still alive in CATS, but they are or will be

positioned and configured differently than they were in KIRIS. The major changes will be the inclusion of a national norm-referenced test, the CTB 5 Survey Edition,[4] and an increase in the salience of student scores.

Maryland

In Maryland, educational reforms began in 1989 when the Governor's Commission on School Performance analyzed the condition of public schools and made recommendations for school improvement. Dubbed "Schools for Success," the reform platform laid out two goals that have had a lasting impact on assessment practices, calling respectively for 70% of students to achieve satisfactory, and 25% to achieve excellent, levels of performance on challenging state assessments by the year 2000. The MSPAP (Maryland School Performance Assessment Program) was created to provide the assessment tools for the accountability system that would gauge progress toward those goals. MSPAP was first administered to grades 3, 5, and 8 in 1991; the revision created for 1992 has remained the standard for new iterations of the instrument since that time. Assessments are given in six areas: reading, writing, language usage, mathematics, science, and social science. All of the tasks in MSPAP are performance tasks, ranging from very short-answer responses to more complex, multistage responses to data, experiences, or text. Each and every task is subjected to extensive review and evaluation, both internal to the state and external, to ensure conceptual and psychometric soundness. Data, both scaled scores and standards-based (percentage achieving various levels), are reported at the school and district levels; individual scores are provided to schools, but schools are warned about the technical and ethical issues that arise from reporting individual scores. There are rewards and sanctions associated with performance on the MSPAP. Schools are rewarded, both financially and publicly, for achieving excellent results. Schools that continue to fall behind and fail to meet their goals may be "taken over by the state" or become "R.E." (reconstitution eligible). While there is no official statewide curriculum, there is a document called the Maryland Learning Outcomes, first approved by the State Board of Education in 1990.

Teacher involvement has been a consistent and distinguishing feature of the MSPAP. Teachers are involved in every aspect of the assessment, from conceptualization to task development to the development of scoring guides to scoring. Maryland educators appear to take pride in this aspect of their work. Both state-level and school-level participants spoke enthusiastically about the staff development opportunities provided by involving teachers so directly in MSPAP; not only did they learn about the nature and benefits of performance assessment, they also learned very directly about the Maryland model. In terms of support, it is interesting to note that when we interviewed Maryland Department of Education (MDE) personnel, they noted that 1998 was the first legislative session in which no bill had been put forward to abolish the MSPAP.

Oregon

The Oregon story is difficult to summarize because the purpose, nature, and direction of the statewide assessment, especially the role of PBAs within the broader system, has changed frequently. The Oregon Statewide Assessment is a "criterion-referenced" assessment founded on the Oregon Content and Performance Standards. The PBA process began formally in 1991 with the passage of the Oregon Educational Act for the 21st Century. Even so, statewide PBA had been going on in writing long before the 1991 bill (referred to by locals as 3565); direct writing assessments were developed, administered, and scored as far back as the 1970s. The early (1991–1995) version of the standards that undergirded the assessment development work was

decidedly interdisciplinary, emphasizing "applied performance integrative outcomes" that spanned curricular boundaries. Pilot performance assessments were developed, implemented, and evaluated during this exploratory period. In 1995, due to new legislation, the development work began to shift toward more discipline-oriented outcomes and assessments. By 1999, the system had evolved into three components: (a) the Statewide Writing and Mathematics Problem-solving Assessment, composed of on-demand performance tasks, (b) the Oregon Statewide Assessment Program, a multiple-choice assessment in reading and literature, mathematics, science, and social studies, and (c) the Certificate of Initial Mastery, the interdisciplinary portfolio of work samples assembled by 10th grade students and scored by teachers in the state.

The Certificate of Initial Mastery (CIM), the one component that retains the original interdisciplinary and application-oriented focus envisioned in 1991, was implemented during the 1998–99 school year, with some reports of resistance from parents and students throughout the state. While the CIM is discussed as if it were a high-stakes assessment, it has, according to our participants, no negative consequences. It is independent of high school graduation; the idea behind it is that students will buy into the CIM because they will see that passing it provides an extra edge when they apply for jobs during or after high school. The other two components went into effect in the 1999–2000 school year, with school- and district-level reporting.

Washington

The PBA story in Washington began with the Commission on Student Learning (CSL), which was created by the Educational Reform Act of 1993 to aid the development of standards, assessment, and accountability systems. Until 1999, when many of the functions of the CSL were assumed by the Office of the Superintendent of Public Instruction, the CSL played the key role in developing and implementing statewide assessments; even with the shift in responsibility, it remains in operation to develop assessments and to advise the Department of Public Instruction on matters of accountability. The Washington Assessment of Student Learning (WASL), which consists entirely of performance tasks, became operational in the spring of 1999 in reading, writing, listening, and mathematics, with science coming on board in the spring of 2000. Standards-based scores are reported at the individual, school, and district levels. Level of accomplishment is reported for individuals, and the percentages of students at various levels of accomplishment are reported for schools and districts. The WASL is closely linked to the Essential Academic Learning Requirements (EALRs), which serve as a broad framework to guide curriculum development at the district and school levels. The accountability system is scheduled to be fully operational by 2003, at which time sanctions will be implemented at the school level. In the interim, schools have been given provisional targets, such as raising their passing rates by 25% over three years. The state mandates the use of norm-referenced examinations in the other grades; currently, that is either the Iowa Test of Basic Skills and the Comprehensive Test of Basic Skills.

LOOKING ACROSS CASES

Context and Purpose

In all five states, performance-based assessment did not stand on its own; in each case, it was initiated as part of a larger school reform or school improvement effort, with an accountability system of rewards and/or sanctions at its core. It is important to recognize that in each case, assessment is viewed as a tool, a means to an end, where the end is the reform of school curriculum and/or the improvement of school

achievement. In some states, particularly those that entered the reform arena early on, assessment led the reform effort, and standards and curriculum frameworks were added (or refined and extended) later. This was true for both Maryland and Kentucky, our two veteran PBA users. In the other three states, all of which are still gearing up for the full implementation of their accountability systems, the development of assessments, standards, and curriculum frameworks tends to proceed in a parallel and highly articulated fashion. Table 1 details key elements in the implementation of assessments and accountability systems.

TABLE 1. IMPLEMENTATION

State	Date			Enabling Legislation
	Legislation	PBA administration required	Accountability implementation	
Kentucky	1990			Kentucky Educational Reform Act
Maryland	1989	1990	1990	Maryland Statewide Performance Assessment System and Sondheim Commission
Oregon	1991		1998	Oregon Educational Act for the 21st Century
Washington	1993		2005	Commission on Student Learning Education Reform and Assessment Law
Delaware	1995	1997		State Content Standards adopted Delaware Student Testing Program

Participants in all five states talked at length about the motivation behind the infusion of PBAs into the statewide assessment and accountability systems, and in each state, similar motives were cited. Most common was the idea that PBA tasks, in contrast to the multiple-choice items more characteristic of norm-referenced assessments, provided a *patently transparent link to curriculum*. Teachers, parents, students, and the general public would be able to "see" how performance on such tasks represented useful things to test and to teach. This link is strong in all five states, but it is especially strong in Oregon, Washington, and Delaware—states in which standards have played a particularly influential role, from day one, in the development of the assessment instruments earmarked for the statewide assessment. In the words of an Oregon public school official, commenting on the Oregon reform movement, which emphasizes close links between standards and assessments:

> I very much believe in the direction of our reform, but within that direction, there's still grounds for dialogue, and they're ongoing. So it's coherent, and it's meaningful, and it's graspable not only for teachers, but for kids.

With these new assessments, there would be little question about the nature of the curriculum and instruction teachers should emphasize in their classrooms. And teachers would gain new insights into student performance in the bargain. As a former teacher in Oregon recounted:

> I had one very bright student, that had done nothing forever, but I knew that he was very bright, and this one question (this one piece) turned that student around. I learned more about (that one student) in one assessment. So if I had any reservations at all about performance assessment, or doing open-ended, or the on-demand questions—I was won over immediately....

A second, and highly related, motive was the connection of PBAs to *real-world performance*, the kind expected in the workplace and in the community. Individuals in all states talked about the need to provide such information to the business community so that it could evaluate the work-readiness of high school graduates. This is an especially important feature of Oregon's system, particularly its Certificate of Initial Mastery for 10th grade students. An explicit driving force of the Oregon system, according to a county-level official, was "that the assessment would be intrinsically engaging because it would mirror real-world situations and expectations that kids (particularly high school kids) will find themselves in." In Maryland, this incentive is central to the PBA movement, appearing in official MSDE materials:

> Performance assessments are designed to measure higher order thinking processes and the application of knowledge and skills to real world situations. For example, a functional test might test a student's ability to do several perimeter problems; a performance assessment would test a student's ability to use perimeter calculations to plot out a garden in the back yard (Maryland State Department of Education,1992).

A third commonly cited motive was *curriculum reform*. There has always been broad-based faith in the potential of PBAs to operate as a lever to promote curriculum reform, especially reform that espouses the importance of reasoning, problem-solving, and applying tools and knowledge gained from school curriculum (e.g., Resnick & Resnick, 1992); participants in all five of our case study states reflected this same optimism. Kentucky, which, along with Vermont and Maryland, entered the PBA arena very early, expressed this goal very directly, with proponents talking openly about a test that required "students to explain how they arrived at their answers," allowing students to "show that they understand the underlying concepts behind facts and how to use that knowledge in the real world" (North Central Regional Laboratory, 1999). In Maryland, both state officials and district personnel talked openly and proudly about building a test "worth teaching to." In Washington, a state official talked about the use of PBAs to leverage a new kind of instruction in the schools, lamenting the fact that "the K–12 system is very much focused on fact, but not on application." A district administrator from Washington, discussing the tension between current instructional practices and the curricular goals implicit in the new assessments, talked about the potential of the new assessments for influencing teaching:

> ...and I'll be honest, not all teachers are there yet. It's like they are teaching one methodology and using an assessment tool that doesn't align with their instructional strategy. So, our staff development program is really keyed to helping teachers make systemic change [that is aligned with the new tests].

In Oregon, the reform agenda was even more transparent. In the words of a county educational official:

> ...the purpose of the assessment was to model applied performance-assessment to push instruction, and to push all of us into a new paradigm (so to speak) of teaching, learning, and assessment.

A fourth motive is related more to the *accountability* system (who is held to what standards) than to the assessment system per se. Accountability systems do not require the use of PBAs; on the contrary, we have many examples of accountability systems driven largely by multiple-choice assessments, save for direct writing

assessments (Michigan, Illinois, and California, to name but a few), but one of the arguments championed by some PBA advocates is that it is precisely in those situations in which the stakes are high when PBAs, with their more transparent ties to curriculum and reform, are essential. In other words, it is when the stakes are high that people are more likely to teach to the test; hence, all the more reason to have tests worth teaching to.

Accountability is a factor in each of these five states, but it is accomplished differently in each. In Kentucky, where accountability is perhaps most central, the PBA effort was originally a part of the KIRIS (now CATS), which was designed to provide results that would help public schools to be "accountable and productive" as a way of raising Kentucky off the bottom of national rankings "in both education and economic development" (Kentucky Department of Education, 1999).[5] In Oregon, the intent is to promote accountability all the way up and down the system, from the individual student (especially through the Certificate of Initial Mastery) to the teacher, to the school, to the district superintendent. In some Oregon districts, administrator salaries are tied to test performance (but teacher salaries are not); accountability at the school level is accomplished with public reporting of scores by school.

These similarities notwithstanding, there are variations in emphasis worth noting among the five cases. For example, the Oregon system is driven more than the others by the connection to real-world performance, a testament to the centrality and impact of the Certificate of Initial Mastery. In Kentucky, school-level accountability, with rewards and sanctions provided on the basis of growth toward the achievement of specific progress goals, is the driving force. In Maryland, Delaware, and Washington, the connection to curriculum reform has emerged as a dominant theme in our interviews.

Development and Scoring

Teachers are involved in the development of assessments in all five states. They are most centrally involved in the review process, serving on committees that are charged with reviewing items and tasks and procedures for administration, scoring, and reporting. They are also involved in the process of selecting anchors for performance tasks and portfolios; given the central role that anchor papers play in "making performance standards real," this responsibility gives teachers a major voice in the process. Maryland and Oregon indicate that they involve teachers in virtually every aspect of the assessment process, while the other three states involve teachers in selected test development activities (see Table 2).

TABLE 2. RESPONSIBILITIES FOR VARIOUS DEVELOPMENT AND SCORING ACTIVITIES

State	Item/Task Development	Test Administration	Scoring	Selection of Anchors	Advice and Review
Delaware	Teachers	Teachers	Contractor	Teachers	Teachers
Kentucky	Teachers	Teachers	Teachers score portfolios	Teachers	Teachers
Maryland	Teachers, with MSDE content chair assistance	Teachers	Teachers, with contractor assistance	Teachers	Teachers
Oregon	Teachers	Teachers	Teachers	Teachers	Teachers
Washington	Commission on Student Learning (CSL) and contractor	Teachers	Teachers (classroom assessments) Contractor (performance assessments)	CSL and contractor	Teachers

Reporting

Conventions for reporting scores are summarized in Table 3. The general pattern is straightforward. Scores from PBA components are generally reported in a standards-based manner, with the percentages of students achieving different levels of accomplishment (usually with a designation about what level constitutes "mastery"). Scores from norm-referenced tests are reported using one or more of the standard conventions for NRTs(e.g., percentiles, NCEs, or the like). In Delaware, where a mixed model is used to determine whether a standard is met, scores for the norm-referenced test will be combined with scores from the PBA component to determine whether students have met particular standards. In Kentucky, school-level scores from the various components are aggregated to create an index that is then compared to a target expectation.

TABLE 3: REPORTING CONVENTIONS ACROSS THE FIVE CASE STUDY STATES

State	Reporting Conventions	Notes
Delaware	Normed data on the Stanford 9 Standards-based data on PBA	Percentage of students achieving various standards is reported for combined SAT 9/PBA scores
Kentucky	A school accountability index based on several weighted factors; 90% of the index from standards-based assessments, about half of which are PBA.	Student scores reported; even more prominent in CATS.
Maryland	PBA	Percentage of students achieving various proficiency levels
Oregon	Norm-based Scale scores Percentage of students achieving various levels of proficiency	Normed scores for the m-c assessments in reading, math, science, and social studies Percentage passing for CIM and PBAs
Washington	PBA (WLAS) at grades 4, 7, and 11 ITBS or CTBS in other grades	Percentage achieving various levels of accomplishment reported on WLAS

Technical Issues

Technical issues are not independent of policy issues. The technical qualities of an assessment affect its credibility. In Kentucky, technical issues of equating and portfolio scoring consistency had gnawed away at the credibility of KIRIS (it is too soon to know about CATS). While reliability has improved steadily over the years, the improvement was not enough to save the math portfolio from being abandoned, most likely because of additional concerns with time and instructional impact. The writing portfolio, however, has survived for two reasons: (a) people perceive the reliability to have improved over time, and (b) compared to the mathematics education community, the portfolio tradition is a more well-established assessment tool within the English education community. In Oregon there is broad-based support for the principles behind the PBA movement, and teachers have been able to agree on scoring and interpretation for performance tasks and work samples. The issue of task generalizability (how many tasks are enough to get a valid and stable estimate of a student's true mastery) plagues the profession, especially now that the CIM is a reality.[6] Only in Maryland do the technical characteristics seem to be regarded as a major asset rather than a liability of performance items. When queried about the success and stability of the exam, which is based entirely on performance items, albeit shorter and less ambitious in scope than in other states, officials in Maryland consistently cited technical standards right alongside building political support. They mention the slow, steady, and deliberate approach they took to task development,

scoring, and the like.[7] They also mention making adjustments to the tasks and items based on the results of technical analyses.

Technical issues interact with level reporting. Sometimes an assessment that yields trustworthy data at an aggregate (e.g., school) level may not yield reliable data for lower-level (e.g., individual) reporting. Maryland officials are keenly aware of the fact that their assessment provides stable estimates for school-level reporting but not for individual-level reporting; hence the warning that they provide about reporting individual scores.

Relationship to Curriculum

Two aspects of the relationship between assessment and curriculum require our attention: (a) the relationship between the statewide PBA and any statewide curriculum documents, and (b) the impact of the state assessment on district and school curricular practices.

ASSESSMENT AND STATE CURRICULUM EFFORTS

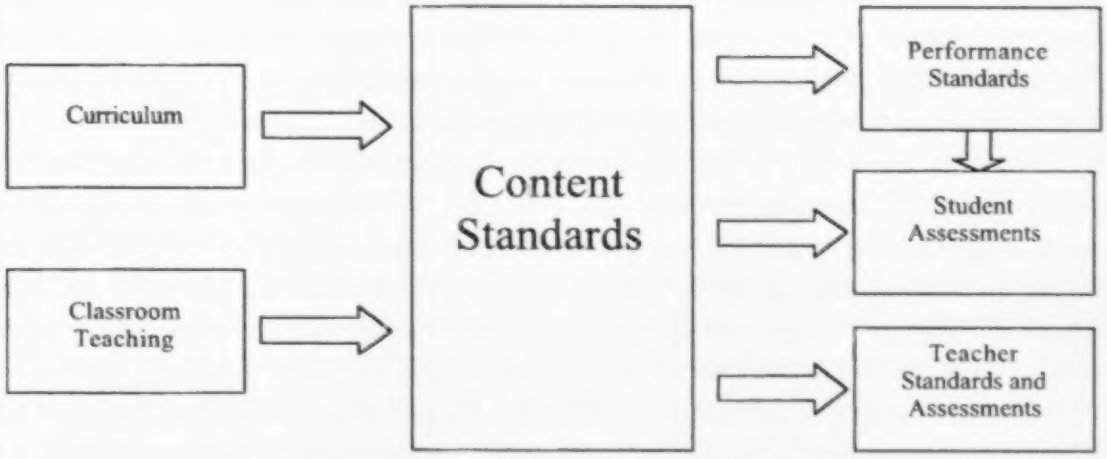
As far as official relationships, depicted in Table 4, state officials in most of the states were careful to point out that there is no official state curriculum, making it clear that curriculum is a district- or school-level prerogative. On the other hand, all five states have in place either curriculum frameworks or content standards, which are intended as very general guidelines for districts and schools to refer to in building their local curriculum. These frameworks also serve as a filter for gauging the appropriateness of items and tasks in the state assessments, although in the case of one state, Kentucky, the assessment preceded the curriculum frameworks in development and implementation.

In two states, Oregon and Delaware, the standards play a particularly important role. In Oregon, a county official refers to the overall effort:

> ...not an assessment driven coalition, but a coalition driven by standards-based education (which means we're driven by standards).... That both instruction and assessment, but also reporting, are all aligned or guided by standards. The assessment in itself is not the driver...and we do have a very defined set of content standards (which convey the expectation of knowledge and skill mastery—that guide the state assessment and presumably guide the work samples).

The Oregon model is the prototypical standards-based reform model, with content standards at the center of the model, driving all other aspects of the system—performance standards, assessment, curriculum, instruction, teacher licensure, and professional development (see Figure 1).

FIGURE 1: A MODEL OF STANDARDS-BASED ALIGNMENT



The Delaware Student Testing Program is built around Delaware's content standards (Delaware Department of Education, 1999). According to a staff member of the department, there is a "one-to-one match" between the curriculum framework and the assessments: "The law says the test 'must' measure the curriculum framework, and that every single item that is on the standards-based portion of the assessment needs to align to a piece of the standards." Apparently, however, some of the standards cannot be assessed on a large-scale assessment and are noted more as "classroom pieces." Thus, all test items and tasks on the state assessment are aligned to standards, but not all the standards are assessed on the state assessment. Specific relationships between state curricular documents and state assessments are detailed in Table 4.[8]

TABLE 4: RELATIONSHIPS BETWEEN STATE CURRICULAR INITIATIVES AND STATE ASSESSMENT

State	State curriculum initiatives	Relationship to assessment
Delaware	Content standards	Assessment must, by law, measure the content standards Scores from m-c assessments figure in any high-stakes, standards-based reporting
Kentucky	Curriculum frameworks	Loose coupling: developed 1 year after KIRIS was in place Since 1997, the Core Content for Assessment has been the basis of test items
Maryland	Curriculum frameworks and content standards (learning outcomes)	Tight link between learning outcomes and MSPAP
Oregon	Academic content and performance standards (1996)	Both state assessment and local curriculum driven by state standards
Washington	Content standards: Essential Academic Learning Requirements	Items and tasks are reviewed for consistency with the EALRs

IMPACT AT THE LOCAL LEVEL

The other side of the relationship between assessment and curriculum is played out not in the creation and alignment of state standards and assessments but in the impact of state assessments on the curriculum that is played out in districts, schools, and classrooms. Here the evidence is mixed across our five case study states. In Kentucky,

the archival evidence is mixed, at least according to one of our participants, who suggested that, "We don't really know how it's [the PBAs] influenced classroom instruction in any systemic way since some research shows it has had a tremendous effect and other research shows it has not."[9] Other Kentucky personnel were concerned with the "sins of omission" that can occur when stakes are attached to a test, even one that promotes authentic performance:

> And I think the state's performance-based assessment has limited what is taught because it goes back to the unkind charge of "teachers are teaching the test." And you better believe it that teachers are teaching to the test because it means their survival and second of all this is the only assessment we have. The negative consequence of this is that other things are not being taught. I also think it has taken a real shot of teacher freedom regarding what a teacher can teach. (A Kentucky school official)

Still other Kentucky participants were more sanguine about the positive influence of the KIRIS assessments, suggesting an overall positive impact on student responsibility for learning and classroom practice.

Our participants from states where the high-stakes accountability feature of the assessment system has not yet been fully implemented—Oregon, Washington, and Delaware—in general had positive perceptions about the impact of the PBA on local curriculum and instruction. A state official in Oregon, which has been easing into the accountability phase of the assessment over several years, offered evidence of the impact of the assessment on student performance:

> We can look at evidences of student work (in mathematics) from two years ago, where we were lucky if a third of the page had writing on it—to this year where the majority of student papers had about two full pages of very well processed work. That is a significant improvement for students focusing on more than one answer, and being able to communicate their process....

And from another Oregon respondent, this account of the activity at the county level:

> ...so we have major curriculum alignment projects...so before we just talk about assessment, we want to make sure the curriculum in schools is providing the teaching/learning opportunity to kids to *prepare* to do well on the assessments.... To provide a system of preparation for kids to reach the standards....

While our participants from Oregon were enthusiastic about the support and positive impact of the PBA effort there, recent data from a broad survey of teachers (Tindal, 1999) suggest that teachers are very concerned about several aspects of the implementation of the Oregon system. In general, the teachers who were surveyed thought that the assessment effort in Oregon had influenced their teaching and that the content of the test aligned with their classroom curriculum. On the negative side, most teachers were mildly negative about the overall impact of the testing system on their teaching and on student learning, and they were extremely concerned about the lack of professional development to do what was being asked of them and the amount of time and energy consumed by the entire enterprise.

In Delaware, which is gearing up for the first standards-based report in 2001 (school-level reporting of the number of students who have achieved the mastery cut-score on the assessment), state officials spoke optimistically about the impact on local curriculum efforts, suggesting that excitement was high about the prospect of demonstrating high levels of mastery in the upcoming assessment, and that the excitement was accompanied by professional development and curriculum alignment. In Washington, where the full force of the assessment system is still a few years off, so is the expectation of its impact on the curriculum. In discussing the impact of the newly developed system of standards and assessments, a state official suggested:

> Well there's good news and there's bad news. When you go into systemic change, you can start with curriculum, or you can start with assessments. And the state has chosen the assessment piece clearly as the driver.... Now the downside of that is that we haven't had an equivalent emphasis...on the curricular piece. So, for example, we have very few districts that would have very strong kinds of (the old word for it was scope and sequence)...so that there's a lack of coherence...[right now].

A local district participant from Washington expressed the view that the key to a positive impact from the state efforts depended entirely on whether resources and staff development would accompany the standards and assessments:

> What's going to change is if I've got ongoing, building-centered staff development that provides me with time to talk and collaborate with other people in my building.... My concern for my staff is we've been doing this with little resources to the point where I worry about my staff.... I mean they're not going to give up tomorrow, but that's that underlying feeling of people hitting the wall, and we don't have enough teachers as it is, what are we going to do then. So somehow something has to kind of move.

There is an overwhelming sense from the people we spoke to in Maryland, both at the state and the local levels, that MSPAP has exerted a positive influence on student performance and classroom practice. As a local participant shared, "The performance assessment and accountability has tremendously changed Maryland. It has put pressure on school and has made positive changes in curriculum and instruction in the State." Participants also talked about the impact of teacher involvement in item development and scoring on professional development in the state, as well as the positive impact on students, who they felt took more responsibility for learning with performance assessments, and the close teacher-student interactions that must occur in preparation for the assessments.

One Maryland interviewee shared a different, but related perspective. In addition to perceived influences on teacher and student behavior attributed to exposure to the MSPAP, it was suggested that at least one aspect of the test administration changed as well:

> At first when the tests came out, they were very secretive and you didn't see the test until the day of the exam when you opened up the test booklet. Then they [MSDE] realized over time that the teacher who was with those children for nine hours, that the teacher is part of the learning environment...and so now the teachers are allowed to look at the test booklets several weeks in advance. They try to make

> sure that the teacher is very comfortable with the test because this will affect how the students are in the test environment.

These observations, when examined across roles (state, district, and school) and states, a few exceptions notwithstanding, support the claim that PBAs can have highly desirable effects on classroom teaching and learning—promoting higher-order thinking, individual responsibility for learning, teacher and student collaboration, focused professional development, and reflection upon the kinds of learning strategies that should be used. While it is difficult, in these analyses, to separate the influence of the PBA component from the accountability systems in which it is embedded, it is highly unlikely that an accountability system without PBAs would have stimulated comments about broadening, rather than narrowing, the impact of assessments on curriculum and teaching.[10]

PROFESSIONAL DEVELOPMENT

Positive effects of PBAs

While professional development was not prompted by a specific question in our interview, it emerged as an issue, both positive and negative, with such regularity that it merits its own treatment. On the positive side, at least one individual (usually more) from each state cited the professional benefits that accrue to teachers who become involved in any aspect of the development and scoring of PBAs. Another benefit cited more than once was the positive and focused impact on teacher learning and teacher-student interaction that can occur when teachers take the state rubrics into their classrooms and analyze them with their students. This benefit is eloquently documented by a county-level participant from Oregon:

> I'm interested in the change from the arms folded attitude to getting lots of students to [look] at the problem.... [Question] "What is it that clicks in a teacher's mind that helps them turn the corner?" [Response] I scored every open-ended question...for a year...in me doing that, and getting the feedback, and learning about their kids...the teachers wanted to take the test themselves ... the teachers realized they were learning a lot more about their students, their thinking process, what they knew, what they didn't know, were they strong in communicating, did they have the content. *Just the value of that turned the corner* [for them].

Another positive professional development venture was implemented in Kentucky. The state hired master teachers, experienced in writing and mathematics, to serve as consultants for districts as they began the process of rolling out PBA practices. Kentucky's emphasis on professional development has continued to be given top priority as it moves away from KIRIS and toward the new CATS as the primary statewide assessment. For instance, the Kentucky Department of Education provides many tools and informational documents on its Web site, including "Common Release Items" and manuals to help teachers develop and incorporate open-response questions into classroom teaching. More recently, the Governor and the Kentucky Education Association appointed a Quality Teacher Taskforce charged with the mission of making sure that teachers are provided with the tools they need to succeed in the classroom.

Concerns about professional development

Professional development, or rather the lack of it, was also viewed as an obstacle to both implementation and any real reform. Recall the concern expressed earlier by the local official from Washington who wondered where the time and resources for the professional development required to implement systemic reform would come from. A state-level participant, also from Washington, likewise recognized the key role of professional development in the assessment-driven curriculum reform effort:

> ...I think we still have a long way to go in turning the curriculum all the way around.... I think that it's more important, the focusing of attention, and giving some direction for curriculum (underscoring the need for a huge staff development/professional development kind of effort)—is probably more important (and I think the assessments had a hand in doing all of those things) and I think those are probably *more important* than the actual test results.

Teacher preparation

At least one individual in each state raised the issue of the role that higher education, through teacher preparation programs, should be playing (and, by implication, was *not* currently playing) in the development of teacher knowledge of PBAs. Interviewees overall believed that there was a real need to inform teacher candidates about PBA use in the classroom and how these assessments relate to statewide reform efforts, but they were not optimistic about getting it done. One interviewee said that courses in teacher prep programs related to PBAs are "pretty much nonexistent," while a Kentucky participant was more hopeful, stating that Kentucky's colleges and universities provide aspiring teachers with "some sort of training related to performance-based assessments." But another interviewee, also from Kentucky, felt that PBA preservice training "needs to occur and it still isn't occurring. There is a lack of change in college-level teaching and it is holding us back." One Maryland state department official captured the disposition attributed in all the states to higher education in his view that while the MSDE certifies the teacher prep programs, "I would say they have been very slow to adopt practices in their programs that are related to the MSPAP." Only in Oregon could we find evidence that a concerted collaborative effort by higher education and state officials had resulted in a substantial and systematic teacher education module on performance-based and curriculum-embedded assessments.

CONSEQUENCES

Rewards and sanctions

Specific rewards and sanctions are part of the accountability system in each of the five case study states, but only two of states, Oregon and Delaware, planned on holding individual students accountable for performance at the time of this study. In the other three states, the school is the basic unit of accountability (although student scores are reported in Washington and are provided in both Kentucky and Maryland), with some provision for district-level accountability in Kentucky.[11] (For the most part, rewards are more common, or at least more prominent, than sanctions, although sanctions exist in all states except Oregon. Table 5 summarizes the nature of the rewards and sanctions employed in each state.)

TABLE 5: CONSEQUENCES (REWARDS AND SANCTIONS) FOR PERFORMANCE ON STATE ASSESSMENTS

State	Nature of the rewards or sanctions
Delaware	Accountability begins in 2001 No takeover, but nonaccreditation is possible Financial rewards for high-performing schools Scholarships for high-performing students
Kentucky	Rewards and sanctions based on meeting individual school improvement goals that require specified progress over time Extremely low-scoring schools are assigned a coach and provided with extra funds Student scores will appear on student transcripts *if they are judged sufficiently valid*
Maryland	Low scores result in classification as a *challenge* school or a *reconstitution eligible* school High scores result in financial rewards
Oregon	CIM will be medium stakes for 10th graders School and district accountability through public reporting and, at local option, administrator salaries
Washington	Accountability begins in 2003, at which time sanctions will kick in There are interim improvement goals (e.g., raise pass rates by 25% per year)

Resistance and support

Assessment is never without controversy, especially when it is part of a high-stakes accountability system, and the PBA components of the assessments in these five states are no exception. That said, these assessments are more remarkable for the support they have attracted than for the resistance they have spawned. In Kentucky and Maryland, the two oldest programs, one might expect resistance to have coalesced by now. To the contrary, Maryland recently experienced its first year without a bill being brought forward in the legislature to dissolve the MSPAP. In 1998, Kentucky reconfigured its accountability system, replacing KIRIS with a newly configured CATS. Some in Kentucky blame the technical inadequacies of KIRIS tests for its demise; others point to the fact that the technical glitches had been fixed before CATS replaced KIRIS and point instead to political opposition to the entire accountability system as the driving force behind CATS. Whether KIRIS or CATS, the business community has remained steadfast in its support for accountability and PBAs throughout both systems. In Delaware, pockets of resistance exist within the teaching profession but not, according to our participants, among the teachers who get involved in the assessment development process; support was noted among all other constituents. No mention of systematic resistance was reported by participants from Oregon, Washington, or Maryland, even when directly queried. In Oregon, several participants mentioned supportive responses from all quarters, although the teachers' union is concerned about the prospects of teacher accountability (Tindal, 1999).

LESSONS LEARNED IN THE CROSS-SITE ANALYSIS

Our conclusions are organized in a set of lessons we learned as we collected and analyzed the data for this study. Our hope is that they exhibit two important characteristics: (a) they follow quite naturally from the analyses presented thus far, and (b) they prove useful to those who work on the PBA agenda in large-scale settings.

1. PAY ATTENTION TO THE TECHNICAL QUALITIES.

A lesson to be learned from the experience of the Maryland PBA is that it is wise to pay attention to the technical qualities of the assessment. While all states have attended to technical issues, in Maryland they were elevated to center stage from the outset. When the state began its PBA development in the late 1980s, Maryland officials made a decision to worry about interjudge reliability, concurrent validity, and generalizability at every step along the way. In doing so, they had to make some compromises with what was, at that time, a more "pure" form of the authentic on-demand task emerging in alternative assessment programs, such as those in the California Learning Assessment System (California Learning Assessment System, 1994) or the first iterations of on-demand tasks in New Standards (Myers & Pearson, 1996). They worked hard on training materials and procedures, as did officials in Kentucky, to improve interjudge reliability each year. They used more, shorter tasks so that they could examine intertask consistency on the way to more conventional estimates of reliability. In retrospect, and if longevity is any index of wisdom, this turns out to have been a wise decision; the Maryland assessment is still with us, albeit in gradually evolving form, while other bolder efforts are now historical artifacts. It is important to note, in this regard, that CLAS assessments and the early New Standards tasks did not fare well on psychometric standards (Linn, DeStefano, Burton, & Hanson, 1995). Conversely, early bad press about technical qualities, ironically at a point when the technical quality (at least of interjudge reliability) was improving, undoubtedly played a role in the repositioning of the PBA components as Kentucky moved from KIRIS to CATS.

2. BUT IT IS MORE THAN *JUST* TECHNICAL.

Technical quality does not guarantee a good or credible assessment. If that were all there was to sustaining support, KIRIS would still be used in Kentucky. Other factors, such as intrusiveness (for teachers and other school personnel as well as for students and their parents), stakes (they can be either too high or too low), and communication (PBAs are easy targets for well-organized disinformation campaigns), no doubt play a major role in the ability of a PBA to maintain its support. The technical qualities are necessary, but not sufficient, to achieve broad acceptance. For broad acceptance, communication appears to be the key. In states where proponents have "taken the test on the road" to bring educators, policymakers, and the general public up to speed, understanding has led to broader support, which brings us to our next lesson.

3. Building Support Is Essential.

Individuals in all five states talked extensively about building a base of support among all key constituencies, especially those with histories of resistance or those who are vulnerable to opposition campaigns. Building support involves a combination of politics, information, and communication. Making sure that those who are likely to resist the effort are brought into the information loop early is a first step; getting early buy-in from these groups is wonderful, but even knowing what sort of opposition to expect is useful. Developing materials that can be shared broadly with the public is also essential, although in the case of PBAs, this is a slippery slope. Because PBAs are unlike what adults experienced in their school careers, they require contextualization and explanation, both of which are difficult in print or electronic communication. Finally, there is no substitute, as our colleagues from Oregon have so ably documented, for getting the word out in public and professional meetings where the PBAs can be explained and where concerns can be expressed and responded to directly. In Maryland, participants cited the broad array of documents available for a variety of audiences, ranging from psychometricians to teachers to parents and the general public, as very helpful in building understanding and support. In Washington, they found that giving the test to the legislature proved invaluable in building understanding of the kinds of standards that can only be captured through PBAs.

4. Establish System Accountability before Student Accountability.

The concern raised by one of our Delaware participants is worth noting. That individual was concerned about new legislation in Delaware making individual student accountability coterminous with the 2001 onset of school- and district-level accountability. The point of the concern is that if system-level accountability precedes individual student accountability, it increases the likelihood that the resources—professional development, curricular practices, a bank of local on-demand tasks, and the like—needed to support students in their quest for higher performance on more challenging tasks will be in place before consequences set in. From all accounts, it appears that the other two states still awaiting the full force of accountability, Washington and Oregon, are trying to take this lesson to heart, although in Oregon, the student accountability assessment embodied in the CIM has preceded school-level implementation by a full year. And, of course, it must be remembered that Kentucky set out on exactly this course almost a decade ago; it is interesting that we see increased pressure for student-level reporting (and accountability) in recent years. As these multiple-level (students, schools, and/or districts) accountability systems play themselves out, legal issues are likely to emerge, as they have in virtually all high-stakes enterprises (Taylor, 2001). While unlikely, it is not implausible to expect rulings that forbid the imposition of individual-level accountability before school and/or district accountability systems have been established and the necessary resources (i.e., the curriculum, materials, and professional development) have been put into place.

5. Nurture Professional Development about and *through* the Assessments.

This is a composite lesson built from the experience of each state. From Kentucky comes the positive experience of establishing a cadre of classroom teachers who can conduct professional development throughout the state; they bring with them an aura of credibility that state personnel and higher education consultants cannot equal. From Maryland and Oregon, and to a lesser extent Washington and Delaware, we take

powerful lessons from the close and intensive involvement of teachers in every step of the assessment process, including task development, rubric development, scoring, and selecting anchors. Teacher involvement in these efforts not only builds ownership and commitment to the process (not to mention its impact on the quality of the assessments themselves), it also builds local capacity; that is, it helps districts and schools to develop their own PBA tasks to use for school- and classroom-based decision-making and to make sure that teachers and students are prepared for the PBA component of the state assessment.

Individuals in several states expressed disappointment in the role of higher education, particularly preservice teacher education. Individuals in several states related their frustration at having to educate new teachers, not only about the state PBA, but about the PBA phenomenon in general, each and every fall. One district-level participant in Washington suggested closer partnerships between teacher education and the public schools they serve to enable new teachers to "hit the ground running" on the PBA front and other equally vexing issues, such as mathematical content knowledge or knowledge of the technical aspects of learning to read. In Oregon, there had been a concerted collaborative effort on the part of higher education, the state education department, and professional organizations to guarantee that preservice teachers get a firm grounding in the use of a wide range of assessment tools, including performance assessments, curriculum-based assessments, and informal classroom assessment. In discussing this issue with state assessment personnel across our five case study states as well as some others, we have become aware of some other efforts to increase the knowledge and awareness of preservice teachers about PBAs and assessment issues more generally. Among the promising initiatives we encountered were co-teaching arrangements in which practitioners and professors worked together to provide a firm grounding in the issues teachers face in everyday teaching (where assessments of various sorts undoubtedly feature prominently) and new permanent certification requirements in which teachers must submit and fulfill an individualized professional development plan linked to both a school improvement plan and state standards for student learning.

6. But It Is More than *Just* Professional Development.

Professional development can provide the knowledge base, and it can even build ownership and commitment among those who participate actively in the assessment process. It can even go a long way to changing deeply held beliefs about the role teachers can play in improving student learning. However, it is unclear whether professional development can change teacher beliefs about the right and responsibility of the state (or the district or any policy body) to hold teachers and schools accountable for student performance. Belief in accountability systems, accompanied by a personal commitment to accepting the responsibility (of improving student achievement) that comes with it, is more a moral and ethical (and political) issue than it is a cognitive and technical matter. Strong, powerful communication, instantiated as persuasion and moral suasion, must accompany technical messages. Persuasion will help, but nothing succeeds like success. As some of the anecdotal evidence about the positive consequences that accrue to students in the classrooms of teachers who embrace performance assessment suggests, when teachers see dramatic changes in student performance and in students' willingness to accept responsibility for their own growth and achievement, they get hooked on performance assessment.

7. If Possible, Do not Send a Test out to Do a Framework's Job.

Because they began the process so early in the national PBA experience, both Maryland and Kentucky used assessments to lead their school reform programs. Standards and curriculum frameworks came later, after the assessments had already begun to affect school curriculum practices. The other three states in our sample, because they have taken a more gradual approach to the development of their school reform efforts, have been able to build assessments alongside the standards or after the standards have been developed and disseminated. If we assume that the assessments themselves are encountering less resistance than they did in the early-entry states (an assumption supported by our interviews), it may be because constituents are better able to focus their concerns on the appropriate reform tool. That is, if a parent or a school board member has a concern about, say, what students are asked to do on a particular task, it may be the content standard that drove the creation of the task, not the task itself, that is the real source of the concern. But if the assessment is the only public document available for public scrutiny, the test, not the underlying standard(s), will bear the brunt of the criticism. Even more important is the contextualizing role that standards and curriculum frameworks can provide for assessment items and tasks. If an assessment artifact can be tied directly to a broadly accepted framework or set of standards, it does not have to stand on its own; the framework provides the assessment artifact with an aura of legitimacy that it might not possess on its own. This is especially important in the case of PBA tasks, which are not likely to strike a familiar chord with adults who experienced very different instructional and assessment artifacts in their school experience.

8. Time May Be a Luxury but It Buys a Lot.

Implicit in many of these lessons is the assumption that time can relieve many of the pressures and sources of resistance that PBAs experience when they enter the policy arena. PBAs are particularly vulnerable to public exposure because they are unfamiliar, time-consuming, and challenging. Time gives space for a more deliberate development process, thus improving the technical quality, and hence the credibility, of our tasks. Time provides for a more elaborate process of communication and public outreach, allowing the assessment to be explained more thoroughly to (and hopefully understood more thoroughly by) a larger set of constituents. The Oregon experience is particularly noteworthy here, and the response of both the profession and the public to an ambitious assessment and accountability system has been accordingly encouraging. Time ensures that assessments will be accompanied by other policy tools, such as frameworks and standards, thus allowing constituents to respond to the assessments as a part of a system rather than as lone artifacts. Time acknowledges the reality that it takes more than a single school year for teachers, other educators, and the general public to acknowledge and accept the fact that new tools and practices are here to stay. In fact, as one of our participants indicated, this whole process is as much about building a new culture of education as it is about changing practices. And time, if accompanied by appropriate resources, gives states, districts, and schools a fighting chance to implement the professional development and to develop the local counterparts of the state assessment tools that make the system work more smoothly and give teachers and students the opportunities they need to put their best foot forward when the assessment becomes a reality.

9. The Most Important Asset of an Assessment System Is the People, Especially the People in the Middle.

As we conducted these interviews, we drew the conclusion that much of the success achieved thus far in these five states must to attributed to the individuals who, at the state, regional, district, and local levels, turn plans for PBAs into realities that play themselves out in schools and classrooms. These are what one of our participants called the people in the middle, those who work between the ends of the continuum. At the one end of the continuum are the legislators and policymakers who enact the mandates and enabling legislation that require these assessment and accountability systems. At the other end are the teachers, students, and parents who must live with the consequences of these assessments. And there are the "people in the middle" —those state-, district-, and local-level personnel who turn the mandates imposed by the policymakers into realities that can be tolerated, if not embraced, by the constituents of the systems. Our conclusion about the important role played by these individuals came not so much from what was said in the interviews as from the sense of dedication and commitment we took away from the interviews. There was a strong sense that this sort of effort has the potential to improve the daily curricular life of students and teachers and to provide parents with better evidence about how their children and our schools are doing. This respect for and reliance on committed educators is probably what was behind the statement, which we heard in one form or another in every site, that "It is technical, for sure, but it is more than just technical."

REFERENCES

Brenner, D., Pearson, P. D., Boyd, J., & Prico, P. (1996, December). *Glendale's New Assessment: The Archival Portfolio*. Charleston, SC: National Reading Conference.

Calfee, R., Pearson, P. D., & Rodriguez, M. (in press). Performance assessment: A concept paper. Unpublished paper, University of California-Riverside.

California Learning Assessment System. (1994). *Elementary performance assessments: Integrated English-language arts illustrative material*. Sacramento: California Department of Education.

Callahan, S. (1997). Tests worth taking? Using portfolios for accountability in Kentucky. *Research in the Teaching of English, 31*(3), 295–336.

Delaware Department of Education (1999). *State summary report 1998 administration*. Dover: Delaware Student Testing Program Assessment and Analysis Group, Assessment and Accountability Branch, Delaware Department of Education.

Fairtest (1995). *Implementing performance assessments: A guide to classroom, school and system reform*. Cambridge, MA: Author.

Hansen, J. (1992). Evaluation: "My portfolio shows who I am." *The Quarterly of the National Writing Project and the Center for the Study of Writing and Literacy, 14*(1), 5–9.

Hoffman, J., Worthy, J., Roser, N., McKool, S., Rutherford, W., & Strecker. S. (1996). Performance assessment in first-grade classrooms: The PALM Model. In D. Leu, C. Kinzer, & K. Hinchman [Eds.], *Literacies for the 21st century* (pp.100–112). Chicago: The National Reading Conference.

Kentucky Department of Education Web site (1999, August). http://www.Kentucky Department of Education.state.ky.us/comm/pubinfo/kystory/.

Koretz, D. M. (1996). Perceived effects of the Kentucky Instructional Results Information System (KIRIS). Santa Monica, CA: Rand Corporation, Institute on Education & Training.

Koretz, D., & Barron, S. I. (1998). The validity of gains in scores on the Kentucky Instructional Results Information System (KIRIS). Santa Monica, CA: Rand Corporation, Institute on Education & Training.

Linn, R. (2000) Assessment and accountability. *Educational Researcher, 29*, 4–15.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*, 15–21.

Linn, R., DeStefano, L., Burton, E, & Hanson, M. (1995). Generalizability of New Standards Project 1993 Pilot Study Tasks in Mathematics. *Applied Measurement in Education*, 9(2), 33–45.

Mabry, L. (1992, December). Twenty years of alternative assessment at a Wisconsin high school. *The School Administrator*, 12–13.

Maryland State Department of Education (1992, March). *MSDE update*. Baltimore: Author.

Meier, D. (1995). *The power of their ideas*. Boston: Beacon Press.

Myers, M., & Pearson, P. D. (1996). Performance assessment and the literacy unit of the New Standards Project. *Assessing Writing, 3* (1), 5–29.

North Central Regional Laboratory Web site (1999, September). www.ncrel.org/ncrel/sdrs/areas/issues/methods/assment/as8lk37.htm.

Pelavin Research Institute. (1996). *Studies of educational reform: Assessment of student performance* (Volume 1: Findings and conclusions). Washington, DC: Author.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement, and instruction* (pp. 37–75). Boston: Kluwer Academic Publishers.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1992). Performance assessments: Political rhetoric and measurement reality. *Educational Researcher, 21,* 22–27.

Simmons, W., & Resnick, L. (1993). Assessment as the catalyst of school reform. *Educational Leadership*, 50(5), 11–15.

Taylor, K. R. (2001). Constitutional law 101: Do high-stakes tests pass or fail? Debra P. v. Turlington. *Principal Leadership* (High School Ed.), *1*(5), 76–78.

Tierney, R. J., Carter, M. A., & Desai, L. E. (1991). *Portfolio assessment in the reading-writing classroom*. Norwood, MA: Christopher-Gordon.

Tindal, G. (1999). *Oregon Education Association: Teacher survey of statewide testing in Oregon*. Eugene: University of Oregon, Educational Research Associates.

Tucker, M. S. (1988). Peter Drucker, knowledge, work, and the structure of schools. *Educational Leadership, 45*(5), 44–46.

Valencia, S. W., Hiebert, E. H., & Afflerbach, P. (Eds.), (1994). *Authentic reading assessment: Practices and possibilities*. Newark, DE: International Reading Association.

Weiss, B. (1994). California's New English-language arts assessment. In S. W. Valencia, E. H. Hiebert, & P. Afflerbach (Eds.), *Authentic reading assessment: Practices and possibilities*. Newark, DE: International Reading Association.

Wiggins, G. (1993). *Assessing student performance. Exploring the purpose and limits of testing*. San Francisco: Jossey-Bass Publishers.

Wixson, K. K., & Dutro, E. (1998). *Standards for primary-grade reading: An analysis of state frameworks*. (Technical Report #3-001). Ann Arbor, MI: Center for the Improvement of Early Reading Achievement, University of Michigan.

RELEVANT STATE-LEVEL RESOURCES

Accountability System Development and Implementation (1999). Template for OFM decision packages. http://www.k12.wa.us/Legis/DP/08.html.

Bergeson, T. (1999, June). State Laws Regarding Education Reform and Assessment in Washington State. Updated. Includes 1999 legislative changes.

Commission on Student Learning (1999, August 16). Information on education reform and accountability: Background information on education reform. http://csl.wednet.edu.

Commission on Student Learning (1999, February). Revised guidelines for inclusion and accommodations for special populations on the Washington Assessment of Student Learning (WASL). Approved for 1999 spring administration for 4th, 7th, and 10th grades. http://csl.wednet.edu.

Delaware Student Testing Program (1999, January 4). *State summary report 1998 administration*. Prepared by the Assessment and Analysis Group, Assessment and Accountability Branch, Delaware Department of Education.

Developing Oregon's content standards (1995–1996). How the standards were developed. (1999, March 28). http://www.ode.state.or.us.

Ensign, G. (1998, May). Washington Assessment of Student Learning: An update. http://csl.wednet.edu.

Kentucky Department of Education (1994). *Measuring up: The Kentucky instructional results information system (KIRIS)*. Frankfort: Author.

Kentucky Department of Education (1994). Kentucky Instructional Results Information System. 1992–93 Technical Report. Frankfort: Author

Kentucky Department of Education's Web site (1999, August). http://www.kde.state.ky.us/comm/pubinfo/kystory/.

Maryland school performance report 1998. State and school systems. Baltimore: Maryland State Department of Education.

Maryland State Department of Education. (1998, April 30). The Maryland School Performance Assessment Program (Revised). **[location?]** Author.

Oregon Department of Education Web site (1999, March 1). State tests: Questions and answers. http://www.ode.state.or.us.

Oregon Department of Education Web site (1999, March 28). Oregon Statewide Assessment Program (1998). http://www.ode.state.or.us.

Oregon Statewide Assessment Results (1998). Press Briefing Packet. Summary of 1998 Statewide Assessment Results. Salem: Oregon Department of Education, Office of Assessment and Evaluation.

Oregon Statewide Writing and Mathematics Problem-Solving Assessment (1999). Administration Manual. Salem: Office of Assessment and Evaluation, Oregon Department of Education.

APPENDIX

INTERVIEW PROTOCOL

April 1, 1999

The Purpose of this Project:

The purpose of this project is to better understand how and why statewide performance-based assessments have developed and evolved as they have. Although we have already collected public documents about performance assessments in your state, we understand that these documents can only be judged as "partial" information and that they are often silent on the question of why and how things developed as they did. This is why we have requested to meet with you and others to further discuss the stories that lie behind your state's performance-based assessments.

Consent Statement:

Your participation in this project is completely voluntary. Your confidentiality is important to us and your input will not be attributed to you individually, but to the project or program you work with. You may discontinue participation in the interview or with the project at any time.

Do you have any questions about your participation in this project or the intent and purpose of the project?

[If interviewee has no questions or concerns...] Do I have your permission to begin the interview now?

Opening Questions:

Let's begin with your view of performance assessment in your state.

What have been your experiences and how do you see the use of performance-based assessments unfolding in [state's name]?

Now let's look at the **Context of [State's Name] Performance Assessments**.

For example: **Mandated versus Internal, Purpose and Audience, Staff Development**

1. When were performance assessments first given in your state, and why were they implemented to begin with?
2. Who first proposed performance-based assessments (i.e., teachers, business community)?
3. In what area(s) was the first performance assessment given (example, language arts for grade 3)?
4. Why was *this* the first performance assessment given?

5. Have the purposes for using performance assessments changed over time? (Where--in what documents--could we learn more about this ?)

6. What is your ballpark estimate of the cost to develop and administer the program?

7. How many staff are dedicated to the performance assessment (full-time equivalents)?

   Have outside consultants been utilized? If so, what role do they play?

8. Are state assessment results used for school-, district-, or student-level accountability?

9. What kind of professional development is provided to teachers or others who administer or score the tests?

10. In your state, is preservice training provided for teacher education candidates regarding the use of performance-based assessments?

11. See attached for ranking purpose.

The Purpose Of The Assessment

According to Pelavin (1996), the purposes of assessment can be summarized into five broad areas: Monitoring, Accountability, Certifying Student Skills, Achieving Better Alignment of Curriculum and Instruction, and Informing and Influencing Curriculum and Instructional Practices.

We want to know your assessment of the relative importance of each of these purposes. To give a sense of your views of their importance, we want you to distribute 100 points across the five categories, giving more points to the more important categories.

PURPOSE	COMMENTS	RANK ORDER POINTS
MONITORING	Student progress toward desired outcomes is a feature of most assessments developed at any level of initiation	
ACCOUNTABILITY (SCHOOLS AND TEACHERS)	For student achievement, formally, through a system of rewards and sanctions, or informally, through mechanisms as reporting school and district performance to media	
CERTIFYING STUDENT SKILLS		
ACHIEVING BETTER ALIGNMENT OF CURRICULUM & INSTRUCTION	The focus of some national and state reform efforts	
INFORMING AND INFLUENCING CURRICULUM, & INSTRUCTIONAL PRACTICES	Frequently cited purpose of assessment reforms	

Let's move into the area of **Curriculum Domain.**

For example: **Conceptualizing a Domain**

1. Do you have a statewide curriculum, curriculum framework, or set of content standards? If so,

   a) How long has it been in place?

   b) How was it developed?

   c) By whom was it developed, and was there public review or input in the development process?

   d) Was it mandated by legislation and funded by the state? Has there been a change over time in the areas covered?

   e) What is the relationship between your statewide curriculum and the performance assessments in place?

2. Do you have performance standards as well as content standards? If so, how are they related?

3. What are your thoughts about how the state's performance-based assessments have influenced local curriculum or classroom instruction?

Let's now discuss the **Development of the Performance Assessments**.

For example: Criteria for Quality, Standards, and Procedures.

1. How were the assessments developed?

2. Has the validity of the program been evaluated, either internally or externally? (Make sure that we have any documents describing this work.)

   If so, have any changes been made to the program as a result?

3. (ask only if they have indicated that performance standards exist in curriculum #3) What is the relationship between state performance standards and the assessments?

4. How was the mix of response formats decided upon (for instance, number of multiple-choice and constructed response questions)?

5. Who has been involved in the development of performance assessments (for example, parents, teachers, students, assessment and curriculum staff, others)? How has this worked?

6. How and by whom are performance items developed? Has this always been the case?

   What do you see as the pros and cons?

7. What psychometric and ethical standards are used to evaluate the program (for instance, the involvement of Bias Review and Technical Advisory Committees)?

Let's consider the **Administration and Scoring** of performance assessments in your state.

For example: Rubrics, Benchmarks, and Progress vs. Accomplishment

1. How and by whom are the actual assessments administered (for example, by teachers, state officials, contractors, others)?

2. How long does it take to administer the tests?

3. How are the items scored, and by whom?

   When are they scored?

   Where are they scored?

   (Make sure we have all relevant reports or training manuals available)

4. Is it mandatory or voluntary? Do all students have to take the tests?

5. What sorts of special accommodations or modifications in test procedures are provided for particular target populations, e.g., students with disabilities (IEP and/or 504), LEP students, or other students with special needs? Please explain.

6 Are any students exempt from taking the tests?

   If so, how is this determined, and by whom?

Let's now talk about the **Reporting System**.

For example: Level of Reporting, Validity, Consequences, Equity, Responsibility

1. How are scores reported for the performance assessments? Are they reported separately or mixed in with any other items?

2. What types of reports are generated? (How are results from students who took the tests with accommodations or modifications treated in aggregate (for example, school- or state-level) reports?)

3. How long does it take to get the results after the test is administered?

4. What was the initial reaction of various groups when the scores for the first performance assessments were given (for example, teachers, parents, legislators, students, others)?

   Has this reaction changed over time as the tests have become more commonplace?

5. What are the consequences of participating in the assessment? For example, what happens to outlier schools, ones that perform either very well or poorly?

6. How heavily are results from performance-based assessments weighted for any sort of accountability reporting?

Other Issues.

1. What kind of recommendations do you have for those in other states interested in implementing performance-based assessments?

2. Has it been your experience that certain areas or subjects seem to be more amenable to performance assessments?

3. What do you see as the pros and cons of performance-based assessments, especially in your state?

4. What would be the ideal large-scale performance-based assessment, and why?

5. Who is supportive of the your State's program:

   (a) Students

   (b) Parents

   (c) Teachers

   (d) Unions

   (e) State Board of Education

   (f) Legislators (the Governor)

   (g) Businesspersons

   Why?

6. Are there any who are not supportive of the program?

7. Are there other key people we should contact to learn more about this whole process, including its history? Perhaps those who were involved but have moved on to other roles or places?

ENDNOTES

[1] We want to publicly acknowledge the benefits of the World Wide Web in accessing state documents. The availability of public documents on the Web has dramatically changed our capacity to include documents in policy analyses such as this one.

[2] Beginning with 1997, assessments were split across grades 4/5, 7/8, and 11/12. Beginning with 1999, some content areas were moved to grade 10 as well.

[3] The mathematics portfolio was discontinued in 1996.

[4] The standardized test will be worth 5 % of the total school index; participants told us that a major reason for including it is to ensure greater agreement between internal and external indices of Kentucky students' performance. It is worth noting that this concern arose originally in 1998 when there was a discrepancy between the pass rates on KIRIS and the classification of Kentucky students on the National Assessment of Educational Progress.

[5] It is important to note that Kentucky's test scores were not at the bottom of the national distribution; they were near the middle. The conjoining of economic development and educational progress suggests an inference at odds with the evidence.

[6] This problem is not unique to Oregon. Task generalizability has surfaced in a number of investigations of performance assessments (e.g., Linn, et al., 1995; Shavelson, Baxter, & Pine, 1992).

[7] Although not extensively mentioned, the question of multiple scoring (examining social studies assessments for both social studies and writing standards) did arise in our conversations as an asset that promoted support.

[8] Our analysis did not reveal whether this same issue about some standards not being assessed on the state test arose in states other than Delaware; however, other research (e.g., Wixson & Dutro, 1998) suggests that such mismatches are common.

[9] The weight of the evidence that we have encountered in both internal and external evaluations of the Kentucky reform movement suggests that it has exerted substantial impact on instruction at the local level. In one account of the impact of the writing portfolio, substantial evidence was presented for a positive effect on both classroom instruction and student learning at the high school level (Callahan, 1997).

[10] See Linn (2000) for a compelling account of the curriculum-narrowing impact that assessments, especially standardized assessments, can have on curriculum and teaching.

[11] In Kentucky, students' scores will be recorded on their high school transcripts provided that they prove sufficiently valid and reliable; districts and schools can and often do share scores with parents and students.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

[X] This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

[ ] This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").